Testimonials on emotions - a multimodal speech analysis

Gaëlle Ferré

Testimonials on emotions – a multimodal speech analysis

Gaëlle Ferré
LLING – Université de Nantes
Chemin de la Censive du Tertre, BP 81227
44312 NANTES Cedex 3
E-mail: gaelle.ferre@univ-nantes.fr

Abstract

The proposal in this pilot study is to present a prosodic and gestural analysis of testimonials on emotions in English and French, the premise being that while recollecting emotions, speakers give some representation of the emotions they convey. They actually shift from neutral to more emotional speech. The corpus described below is a series of podcasts in which ordinary people from different countries have been video recorded giving their opinion on issues that involve emotional speech. Three emotions have been considered: cold anger, sadness and happiness, and in terms of prosody, the paper concentrates on the intonation contours used by the speakers rather than on pitch span and F0 level. The annotation of gestures made with ELAN is based on a modified version of the MUMIN Coding Scheme (Allwood et al., 2005) and the conclusion of the paper is that speech is felt to be more emotional when a set of prosodic and gestural parameters is modified from more neutral speech, e.g. smiles or laughs do not in themselves convey happiness, but must be combined with other gestural and prosodic features to do so.

1. Introduction

This paper proposes a multimodal analysis of emotional speech compared to more neutral parts of discourse in podcast films recorded by the association "GoodPlanet" for a public exhibition directed by photographer Yann Arthus-Bertrand. The corpus description (see section 2) addresses the issue of the use of podcast recordings for research purposes. Among the films recorded by the association, I chose to analyze a collection of testimonies which where semantically linked with emotions, the underlying principle being that while speaking of their emotions, the speakers show a bit of these emotions in their expression. According to Ekman (1999:50) expressions "are part of an emotion; they are a sign that an emotion is occurring". This principle illustrates what Caffi & Janney, quoted in Plantin (2003:9), call emotional communication, i.e. "a type of spontaneous, unintentional leakage or bursting out of emotion in speech". And this is exactly what occurs in the videos, where speakers narrate some emotional experience to the public in a rather neutral tone, but at times, their emotions burst out in their intonation and facial expression. In section 4, I present a qualitative analysis of some salient prosodic and gestural features related to cold anger, sadness and happiness.

2. Corpus description

The corpus I worked on is a series of podcasts made available online by the association "GoodPlanet" headed by photographer Yann Arthus-Bertrand. The aim of the association's project "6 billion others" is to collect a series of testimonials from people around the world on a vast quantity of topics. Ordinary people from different parts of the world were asked to give their opinion on questions such as "what do you fear most?", "what has been the biggest joy in your life?, "on what occasion did you cry?", "what makes you particularly angry?" or "have you ever felt discrimination?". They were video-taped by a team of filmmakers of the association and the videos were edited into clips, the utterances said being inserted as a French subtitle into the clip. Each clip shows up to five or six people speaking different languages and expressing an opinion on a common topic, and lasts about 5 minutes. The project has been running for five years now and the final aim is to show the videos (more than 5000 interviews filmed in about 75 countries) in a free exhibition to be held in the Grand Palais in Paris in January and February 2009. The association has given consent for the video-clips to be used in a research project on emotions, provided the persons are treated with dignity and respect. This doesn't completely answer the issue on ethics, since such research work was not initially covered by the association's project, but the participants were recorded of their free will and accepted that the recordings be used in a public exhibition and displayed on the internet. No judgement will be made in this paper on the participants' physical appearance or on the relevance of their speech.

The major difficulty while working on podcasts is the issue of the naturalness of the data. The clips I worked on have been edited by professional filmmakers who obviously cut the sequences of hesitations and false starts always present in spontaneous data, but the resulting films are very close to natural data. It just means, in terms of prosody, that the clips can't be used for a study of speech rhythm, silent pauses and hesitation discourse markers, which is not the purpose of this paper. Another drawback of the clips is that the persons have been filmed during their testimonials but no unemotional speech of the same speakers has been recorded for comparison and although each clip contains testimonials on the same topic, speakers do not say exactly the same thing. Consequently, the object of the paper will be to study the shifts between neutral and emotional speech within each testimonial rather than comparing emotional speech between speakers.

When using podcasts, the major drawback lies in the often poor quality of the recordings. Most of the time, the recordings use a compression rate which is too low for a good image quality in terms of pixel number, a quality which is nevertheless necessary to allow even a manual

detection of fine facial details. This is not the case of the *GoodPlanet* podcasts which are of a high quality despite compression. The clips were edited in m4v, but I changed the codec into MPEG1 to be able to read the clips in ELAN (see section 8). Another drawback of podcasts is that speakers are usually filmed in action, with numerous zooms and close-ups which make it difficult for the analyst to annotate the movements of a particular part of the body, not always visible on the video. The testimonies of *GoodPlanet* are all filmed in the same condition, with an extreme close-up on the speaker's face. This is particularly convenient for the annotation of fine details of the face, although it means, of course, that other parts of the speakers' bodies are not visible on the screen. At last, a prosodic analysis of speech requires an extremely good quality of the sound of the recordings, and this is far from being the case in most podcasts. In the vast majority of podcasts, the sound-track is full of background noise, either music added to the initial recording, or surrounding noise during the recording which prevents any serious detection of prosodic parameters such as F0 or intensity. Another problem lies in the fact that in many recordings, several people speak in overlap and are recorded with a single microphone which also hinders the detection of prosodic parameters. *GoodPlanet* podcasts present none of these faults since the speakers are recorded professionally with no overlapping speech and background music or noise. The sound tracks of the clips are of a good quality and allow prosodic treatment of the data.

3. Annotation of the corpus

3.1 Phonology and prosody

The first step in the annotation process consisted in a prosodic annotation of about 15 mn of the corpus collection with PRAAT (see section 8). The sound files, extracted from the videos, were first converted into wave files to be readable in PRAAT and transcribed into current orthography in what could be termed intuitive prosodic groups. The translations made by the filmmakers could not be used since they didn't correspond to the exact words uttered by the speakers and were not aligned with the sound signal. From this initial transcription, words, syllables and phonemes were then annotated using the SAMPA convention of annotation, entirely manually for English and with the help of EasyAlign for French (see section 8). I then noted focal accents on a separate tier and on yet another tier, prosodic units (intermediate vs. intonational phrases, see Cruttenden, 1997:59-60) were noted as well as the general pitch contour on each phrase, using the contours described in Cruttenden (1997): high and low fall, high and low rise, rise-fall, fall-rise, and flat. In the cases where PRAAT displayed F0 detection errors, I used the Prosogram (see section 8) to obtain a better representation of the contour. The phonemic and prosodic annotation is shown in Figure 1 below:

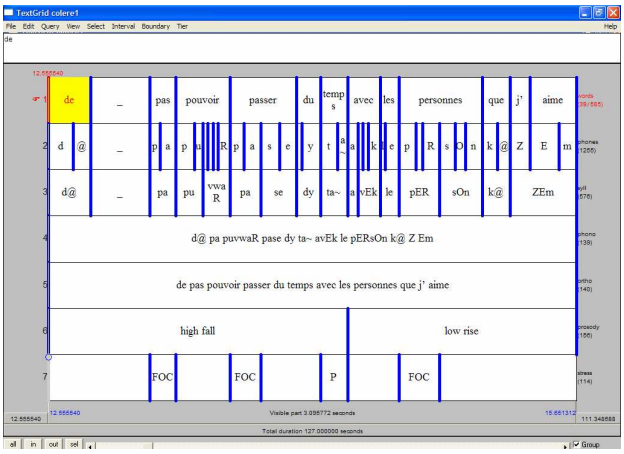

Figure 1: Textgrid of the orthographic, phonemic and prosodic annotation of the corpus in PRAAT.

3.2 Gestures

These annotations were then imported in ELAN, a tool for the annotation of video files (see section 8). For the annotation of gestures, I used an adapted version of the MUMIN coding scheme, fully described in Bertrand et al. (2007). The scheme was developed to describe gestures made in interactional data, which is quite poor in emotional load (Bouchard, 2000). Yet, the description of gestures using this scheme is extremely precise and a tier has been devoted to the annotation of emotions/attitude. I compared this scheme with the one recommended by the Network of Excellence HUMAINE and the annotation types of the two schemes are very close. Since the films were close-ups of the speakers' faces, I only annotated facial movements in a first step: gaze direction (front, left, right, up, down), head direction and head movements (shake, nod, tilt, beat...), eyebrow movements (frown, raise), mouth movements (smile, laugh...). In the file concerning sadness, I also added a track to annotate eye moisture (wet eyes) which was not initially thought of in the scheme, since this parameter seemed to play a role when the speaker displayed this emotion.

At last, in each file, I annotated the parts where speech seemed to be more loaded with emotion, simply stating the emotion (cold anger, sadness, happiness). This was made rather intuitively and by only one annotator for the moment. It is clear that in the future, several annotators will be needed for this annotation which is more subjective than the mere annotation of gestures, and their agreement checked by Kappa measurements.

Correlations of the different parameters were made with the help of the search function of ELAN.

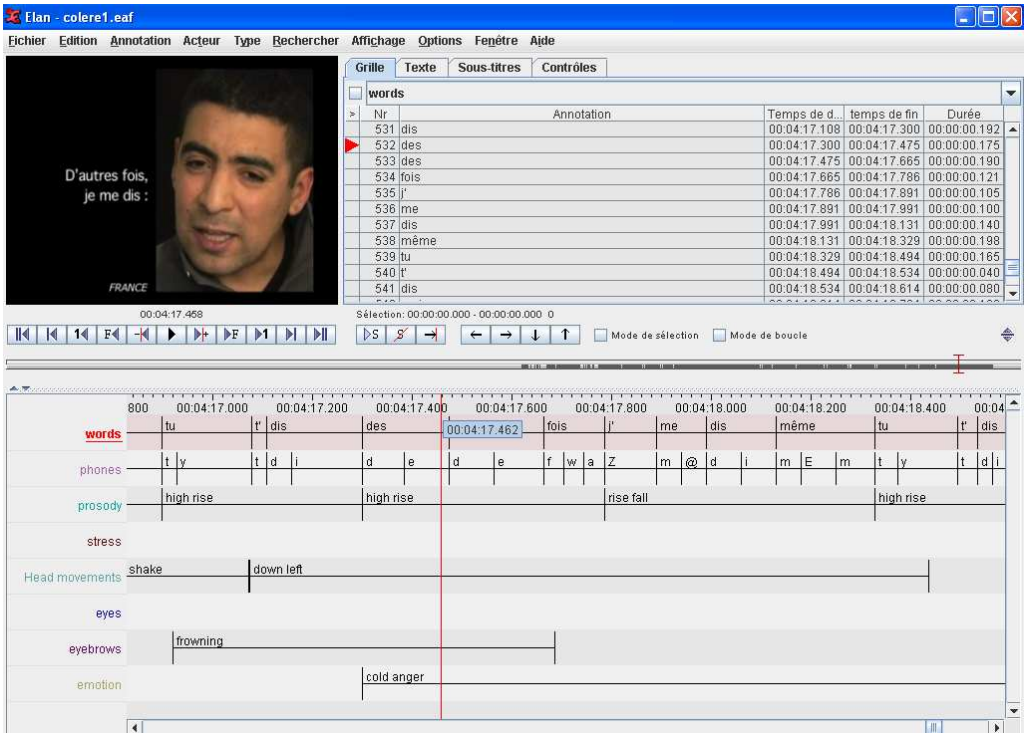

Figure 2: Snapshot of the annotation board in ELAN showing orthographic, phonemic, prosodic and gesture tracks.

4. Results and discussion

Although the files were too short to present any statistical analysis here, the results obtained are nevertheless interesting in a qualitative approach which may constitute the basis for further work.

4.1 Prosody

As stated earlier in this paper, the edited nature of the files does not allow work on rhythm (filled and silent pauses as well as speech rate) since parts of the file have probably been deleted and we do not have access to the original recordings. Yet, much can be done on pitch and stress. Previous studies have mostly described the links between emotional speech and speech span as well as F0 average height. Working on Japanese and French, Yamasaki (2004), for instance, found that positive emotions were rather perceived with rising F0, high average F0 being associated with gaiety. On the contrary, negative emotions were rather perceived when the F0 was falling, low average F0 being associated with sadness. She didn't however annotate contours from a phonological perspective. Devillers and Vasilescu (2004) studied the span and height of the intonation contour and found that in two negative emotions, fear and anger, the amplitude of the parameters depended upon subjects, but tended to be larger in emotional than in neutral speech. Both studies agree on the important weight of lexical/semantic content in the perception of emotional speech.

In this study, I concentrated on phonological F0 contours, comparing the distribution of contours for each speaker on the total testimony on one hand, and in correlation with emotions on the other hand. Results showed that there is a larger proportion of rising-falling contours in sadness as in the total testimony for this speaker. Happiness, on the contrary, was more correlated with low rising contours. This is in agreement with the findings of Yamasaki (op.cit.) associating lower intonation to negative emotions and higher intonation to positive ones. Looking at cold anger, however, an emotion which could be classified as negative, one finds that there is a higher proportion of high rises and rising-falling contours which do not correspond to Yamasaki's results. Yet, they confirm Devillers and Vasilescu's results (op. cit.) stating that the pitch span is larger in emotional speech, since I didn't find any high rising contours in declarative utterances in other contexts.

What makes sense however is that emotional speech implies a higher involvement of the speaker in his discourse and this is shown prosodically speaking with the use of a higher proportion of modulated tones such as the rising-falling contour. According to Morel & Danon-Boileau (1998), rising contours show a greater appeal to the co-participant in conversational data on the part of the speaker. In this context, which is not a conversation, the expectancy of empathy is nevertheless still present since the speakers may be said to convey a message to the public, represented by the eye of the camera. Emotional speech then can be described as conveying more involvement on the part of the speaker who also expects a greater involvement on the part of the addressee. Bagou (2001) and Plantin (2003) find that emotional involvement is regularly associated with emphasis in spontaneous speech. I also found that there was a high quantity of focal accents in the files I treated,

especially in the one showing cold anger. The focal accent is mainly realized through lengthening of word initial onset consonants. Yet, speakers in this file switched from rather neutral speech to speech showing cold anger and the proportion of focal accents was not higher in the speech parts perceived as more emotional. In other words, I could not find a direct link between prosodic emphasis and emotional speech, which shows that other parameters are needed for speech to be perceived as emotional. This is even reinforced by the gestural analysis (*Cf.* Analysis of "beats") in the next section.

4.2 Gestures

To follow my thread of thinking, the gesture annotation of the corpus showed that in the cold anger file, most focal accents were accompanied by a head beat of speakers (e.g. a rhythmic downward rapid movement of the head produced on accented syllables). These beats were not either associated with the perceived emotional parts of the file. This shows that a strong focal accent, marked both prosodically and gesturally, is not sufficient for speech to be perceived as loaded in emotional weight, although the utterance is not perceived as neutral, but as emphatic. In my opinion, emphasis and emotion are not necessarily linked.

Other gestures and physical properties were however typically associated with the three emotions under study. What appears immediately in Table 1 is that each gestural parameter allows a distinction between two emotions. For instance, head movements allow a distinction between sadness and happiness (shakes being associated with sadness and nods with happiness), whereas no particular head movement is found in correlation with cold anger.

	Gesture	Cold anger	Sadness	Happi-ness
Eyebrows	Raising		X	
	Frowning	X		
Head	Shakes		X	
	Nods			X
Gaze	Away		X	
	Front			X
Eyes	Narrowed eyes	X		
	Wet eyes		X	
Mouth	Laughs	X		X
	Smiles			X

Table 1: Relevant gestures and physical properties related to the three emotions.

Table 1 corroborates the results of Smith & Scott (1997), partially based on Darwin (1981), especially concerning eyebrow movements (raised eyebrows conveying sadness as opposed to frowns which convey cold anger).

I also found like Smith & Scott that laughs and smiles are associated with happiness.

Figure 3: Still pictures of speakers showing cold anger (left) and happiness (right).

What is different though is that laughs are also met in cold anger in this corpus, which is not however surprising. As stated in Bertrand et al. (2000), laughter reveals a distance of the speaker from his feelings. Laughter can then be met in other feelings than happiness, such as in cold anger and may be induced by the fact that the speaker suddenly realizes his greater involvement in a negative emotion from which he needs to take a distance. This is directly linked as well to the question of politeness evoked in Kerbrat-Orecchioni (2000:51) for whom « la civilité n'admet pas les manifestations émotionnelles intempestives et incontrôlées » and « la politesse et les rites sociaux ont précisément pour fonction principale de canaliser le flux affectif, de juguler les débordements émotionnels (...) ». This is probably also why speakers conveying sadness tend to look away from the camera when the emotion is too strong, whereas in happiness, they look straight at the camera, this latter emotion being a positive one. No regularity in gaze direction could be observed when the emotion conveyed is cold anger. At last, as far as eyes are concerned, this work corroborates Smith & Scott's findings (op. cit.) in which anger was associated with narrowed eyes. So far, the results for cold anger and happiness also corroborate the studies made by Cosnier & Huyghues-Despointes (2000) on the cognitive representation of emotions, a study in which they find that the mental representation of anger triggers movements in the subjects' eyes and eyebrows, whereas the representation of happiness triggers mouth movements. Smith & Scott did not test eye moisture, but in the file I worked on, it was obvious that sadness was perceived in a stronger way with greater eye moisture of the speaker.

Figure 4: Eye moisture in the perception of sadness.

This parameter is however not as reliable as the other ones though, since eye moisture on a video file may depend much on the eye colour of the speakers, as well as on the lighting used during the recording.

5. Conclusion

As a conclusion, I may say that this paper addressed several issues currently under discussion in the research community on emotions: firstly, concerning the data which lacks cruelly, it was shown that some podcast recordings may be used although work on such recordings cannot answer all the questions raised (but is this not the case of any corpus?). The corpus of the project "6 billion others" is of a quality which allows a prosodic and gestural treatment and I showed how speakers, through emotional semantic content, switch from rather neutral narration of their feelings to more emotionally involved speech.

The emotions conveyed were cold anger, sadness and happiness, which I understand as attenuated forms of anger, deep sorrow and joy. They were expressed through certain intonation contours: high rises and rising-falling contours for cold anger, rise falls for sadness and low rises for happiness. These preferred contours show more involvement on the part of the speaker who also seeks some empathy from the addressee. The speaker's involvement is however not linked to the presence of focal accents in this corpus.

In terms of gestures, emotions are conveyed with a set of different parameters: cold anger is associated with eyebrow frown, narrowed eyes and even laughter at times; sadness is associated with raised eyebrows, averted gaze, eye moisture and head shakes; happiness is associated with head nods, direct gaze, smiles and laughter. What both the prosodic and gestural analyses show is that prosody and gestures are not redundant in emotional speech, as already stated by Aubergé (2002), and that both participate in the perception of emotion. It also shows that analyses of video films are useful to complement studies based on pictures, since a laughing face in itself does not necessarily convey happiness, for instance. Yet, a still photograph of a laughing face would nevertheless be interpreted as a happy face without the context of the film and other parameters as already pointed out by Beavin Bavelas & Chovil (1997).

The limits of this study are also important. The quantity of data treated was quite small and an annotation of more recordings would be needed first to confirm the results of this pilot study, then to be really able to compare the results obtained on English and French which were treated here as equivalent.

6. Acknowledgements

I am particularly indebted to A.-L. Charriot and F. Gilard, producers of the project "6 billion others", for allowing me to use the recordings of the association "GoodPlanet", headed by photographer Yann Arthus-Bertrand. I hope the 2008 exhibition in Paris is a success and that useful work can be done in linguistics using their material. Any error in the description of the corpus or the aim of the association is mine.

7. References

Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navarretta, C. & Paggio, P. (2005) The MUMIN Multimodal Coding Scheme, *NorFA yearbook 2005.* http://www.ling.gu.se/~jens/publications/B%20files/B70.pdf

Arthus-Bertrand, Y. (2003-2008). 6 billion others, a project of the Association "GoodPlanet", A.-L. Charriot - Directrice de Production & F. Gilard - Chargé de production, http://www.goodplanet.org/ (last access: 06-02-2008).

Aubergé, V. (2002). Prosodie et émotion. In *Actes des Deuxièmes Assises Nationales du GdR* I3, pp. 263—273,www.irit.fr/GDR-I3/fichiers/assises2002/papers/15-ProsodieEtEmotion.pdf (last access: 07-04-2008).

Bagou, O. (2001). Validation perceptive et réalisations acoustiques de l'implication emphatique dans la narration orale spontanée. *Cahiers de Linguistique Française*, 23, pp. 39--59.

Beavin Bavelas, J., Chovil, N. (1997). Faces in dialogue. In J. A. Russel & J.M. Fernandez-Dols (Eds.), *The Psychology of Facial Expression*. Cambridge, CUP, pp. 334--346.

Bertrand, R., Blache, P., Espesser, R., Ferré, G.; Meunier, C, Priego-Valverde, B., Rauzy, S. (2007). Le CID: Corpus of Interactional Data -protocoles, conventions, annotations-, *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence (TIPA)*, 25, pp. 25-55.

Bertrand, R., Matsangos, A., Périchon, B., Vion, R. (2000). L'observation et l'analyse des affects dans l'interaction. In M. Doury, V. Traverso, C. Plantin (Eds.), *Les émotions dans les interactions*. Lyon, Presses Universitaires de Lyon, pp. 169--182.

Bouchard, R. (2000). M'enfin !!! Des "petits mots" pour les "petites" émotions ? In M. Doury, V. Traverso, C. Plantin (Eds.), *Les émotions dans les interactions*. Lyon, Presses Universitaires de Lyon, pp. 223--238.

Chung, S.-J. (2001). Efficiency of the final part of the utterance in the communication of emotion. In C. Cavé, I. Guaïtella, S. Santi (Eds.), *Oralité et Gestualité (ORAGE) : "Interactions et comportements multimodaux dans la communication"*, Paris, L'Harmattan, pp. 183--189.

Cosnier, J., Huyghues-Despointes, S. (2000). Les mimiques du créateur, ou l'auto-référence des représentations affectives. In M. Doury, V. Traverso, C. Plantin (Eds.), *Les émotions dans les interactions*. Lyon, Presses Universitaires de Lyon, pp. 157--168.

Cruttenden, A. (1997). *Intonation*. Cambridge, CUP, second edition.

Darwin, C. (1981). *L'expression des émotions chez l'homme et les animaux*. Translated from English by S. Pozzi & R. Benoît. Bruxelles, Editions Complexe.

Devillers, L., Vasilescu, I. (2004) Détection des émotions à partir d'indices lexicaux, dialogiques et prosodiques dans le dialogue oral. In B. Bel & I. Marlien (Eds.), *XXVèmes Journées d'Étude sur la Parole*, AFCP, pp. 169--172.

Ekman, P. (1999). Emotional and conversational nonverbal signals. In L. S. Messing & R. Campbell (Eds.), *Gesture, Speech and Sign*. New York, Oxford University Press, pp. 45--55.

Kerbrat-Orecchioni, C. (2000). Quelle place pour les émotions dans la linguistique du XXe siècle ? Remarques et aperçus. In M. Doury, V. Traverso, C. Plantin (Eds.), *Les émotions dans les interactions*. Lyon, Presses Universitaires de Lyon, pp. 33--74.

Morel, M.-A., Danon-Boileau, L. (1998). *Grammaire de l'intonation : L'exemple du français*. Paris, Ophrys.

Network of Excellence HUMAINE. http://emotion-research.net/ (last access: 07-04-2008).

Plantin, C. (2003). Structures verbales de l'émotion parlée et de la parole émue. In J.-M. Colletta & A. Tcherkassof (Eds.), *Les émotions. Cognition, langage et développement*, Liège, Mardaga, pp. 97--130.

Smith, C.A., Scott H.S. (1997). A Componential Approach to the meaning of facial expressions. In J. A. Russel & J.M. Fernandez-Dols (Eds.), *The Psychology of Facial Expression*, Cambridge, CUP, pp. 229--254.

Yamasaki, H. (2004). Perception des émotions "positives" et "négatives" chez les auditeurs français et japonais à travers le contour de F0. In B. Bel & I. Marlien (Eds.), *XXVèmes Journées d'Étude sur la Parole*, AFCP, pp. 465--468.

8. Tools

Praat (Boersma P. and D. Weenick): A system for doing phonetics by computer. http://www.praat.org

EasyAlign (J.-P. Goldman): http://latlcui.unige.ch/phonetique/

Elan (H. Sloetjes): http://www.lat-mpi.eu/tools/elan/

Prosogram (P. Mertens): http://bach.arts.kuleuven.be/pmertens/prosogram/